# Markov Chains and Algorithmic Applications: WEEK 10

*Reminder.* We would like to sample from a distribution $\pi = (\pi_i,\, i \in S)$ on state space $S$. One option for this is the Metropolis-Hastings algorithm:

1. Consider a base chain on $S$ with transition probabilities $\psi_{ij}$
   (irreducible, aperiodic and such that $\psi_{ij} > 0$ if and only if $\psi_{ji} > 0$)
2. Define acceptance probabilities $a_{ij} = \min\left(1, \dfrac{\pi_j\, \psi_{ji}}{\pi_i\, \psi_{ij}}\right)$
3. Define then
$$p_{ij} = \begin{cases} a_{ij}\, \psi_{ij} & \text{if } j \neq i \\ \psi_{ii} + \sum_{k \in S \setminus i} \psi_{ik}(1 - a_{ik}) & \text{if } j = i \end{cases}$$
4. The Markov chain on $S$ with transition probabilities $p_{ij}$ is such that $p_{ij}(n) \underset{n\to\infty}{\longrightarrow} \pi_j$ and detailed balance holds. Running then the Markov chain from an arbitrary initial state $i \in S$ for a sufficiently large amount of time (so that $p_{ij}(n)$ is indeed close to $\pi_j$ for all $j \in S$) is a way to (approximately) sample from $\pi$.

## 1 Application: optimization of a function

Let $f : \mathbb{Z} \to \mathbb{R}$ be a function to be minimized, which is assumed to be bounded from below and such that $\lim_{i\to\pm\infty} f(i) = +\infty$ (so that at least one global minimum exists).

*Problem:* If the function $f$ is complicated and has many local minima, then (greedy) algorithms usually fail to converge to a global minimum[1].

*Our aim:* to sample from the distribution

$$\pi_\infty(i) = \frac{\mathbb{1}_{\{i \text{ is a global minimum of } f\}}}{Z_\infty = \sharp \text{ global minima of } f}, \quad i \in \mathbb{Z}$$

Sampling from $\pi_\infty$ is a difficult task because

1. we have to compute $Z_\infty$, and
2. the global minima may be very isolated on the state space, hence checking the neighborhood of $i$ is not sufficient to compute $\mathbb{1}_{\{i \text{ is a global minimum of } f\}}$.

Instead, we will sample from the distribution $\pi_\beta$:

$$\pi_\beta(i) = \frac{e^{-\beta f(i)}}{Z_\beta}, \quad i \in \mathbb{Z}$$

where $\beta > 0$ is a fixed parameter and $Z_\beta = \sum_{i\in\mathbb{Z}} e^{-\beta f(i)}$ is the normalization constant (that might still be difficult to compute). The idea is that as $\beta$ increases, distribution $\pi_\beta$ concentrates around the global minima of $f$, hence $\pi_\beta \overset{\beta\to\infty}{\longrightarrow} \pi_\infty$.

To avoid computing $Z_\beta$, we will use the Metropolis-Hastings algorithm to construct a Markov chain having $\pi_\beta$ as its stationary distribution:

[1] This typically also happens when $\mathbb{Z}$, the domain of the function, is replaced by a finite but high-dimensional domain.

1. We choose a simple irreducible base chain (such that $\psi_{ij} > 0$ iff $\psi_{ji} > 0$), the symmetric random walk on $\mathbb{Z}$: $\psi_{i,i\pm1} = \frac{1}{2}$ (remember that this chain has no stationary distribution, yet this does not influence the algorithm in any way).

2. The acceptance probabilities are

$$a_{ij} = \min\left(1, \frac{\pi_j}{\pi_i}\right) = \begin{cases} \min\left(1, e^{-\beta(f(j)-f(i))}\right) & j = i \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

   In words, we always accept a transition to a state with a lower value of $f$, but we still accept some non-favorable transitions to avoid getting stuck in a local minimum.

3. The constructed chain having transition probabilities

$$p_{ij} = \begin{cases} \psi_{ij} a_{ij} & j \neq i, \\ 1 - \sum_{k \neq i} \psi_{ik} a_{ik} & j = i, \end{cases}$$

   is such that $p_{ij}(n) \overset{n\to\infty}{\longrightarrow} \pi_{\beta_j} \quad \forall j \in S$.

## 1.1 How to choose $\beta$ ?

Let us give a ballpark estimate to choose $\beta$ correctly. Note that this is just a qualitative idea which can only serve as a first guide when these ideas are applied to specific problems. To choose $\beta$, we decide that we want to spend a $1-\epsilon$ fraction of time in global minima. Recall that $\pi_i$ is the average fraction of time that the chain spends in state $i$ when it has reached the stationary distribution. Thus we set

$$1 - \epsilon \approx \sum_{i \text{ global minimum}} \pi_\beta(i)$$

Let $f_0 = \min_{i\in\mathbb{Z}} f(i)$ be the global minimum and $f_1 = \min_{i\in\mathbb{Z}, f(i)\neq f_0} f(i)$, $f_2 = \min_{i\in\mathbb{Z}, f(i)\neq f_0, f_1} f(i)$, ... be the local minima. Let $N_0$, $N_1$, $N_2$, ... be the number of points were the minima $f_0$, $f_1$, $f_2$, ... are reached. We have

$$\sum_{i \text{ global minimum}} \pi_\beta(i) = \frac{N_0 e^{-\beta f_0}}{Z} \quad \text{and} \quad Z = \sum_{i\in\mathbb{Z}} e^{-\beta f(i)} = \sum_{k\geq 0} N_k e^{-\beta f_k} \approx N_0 e^{-\beta f_0} + N_1 e^{-\beta f_1}$$

(as we think of $\beta$ being reasonably large and $f_0 < f_1 < f_2 < \ldots$). Therefore:

$$\sum_{i \text{ global minimum}} \pi_\beta(i) \approx \frac{N_0 e^{-\beta f_0}}{N_0 e^{-\beta f_0} + N_1 e^{-\beta f_1}} = \frac{1}{1 + \frac{N_1}{N_0} e^{-\beta(f_1 - f_0)}} \approx 1 - \frac{N_1}{N_0} e^{-\beta(f_1 - f_0)}$$

Remembering that we want this term to be approximately equal to $1-\epsilon$, we obtain the following rough estimate for $\beta$:

$$\beta \approx \frac{1}{f_1 - f_0} \log\left(\frac{N_1}{\epsilon N_0}\right)$$

## 1.2 In practice: simulated annealing

The choice of $\beta$ can influence the output of the Metropolis algorithm significantly:

- If we choose $\beta$ large, then $\pi_\beta$ is close to $\pi_\infty$, but the chain produced by the algorithm converges very slowly due to the high probability given to self-loops. In essence, the chain can almost become reducible.
- If we choose $\beta$ small, then the chain produced by the algorithm converges quickly to the stationary distribution $\pi_\beta$ at the cost of potentially being very far from $\pi_\infty$.

The ideal solution would be to combine the best of both worlds, similarly to creating certain alloys: simply mixing the metals at high temperature then immediately bringing the system down to room temperature does not give the alloy the desired properties. Instead, the temperature should be decreased at a slow speed for the metals to bond appropriately.

Consider $\beta$ as representing an inverse temperature. Then the annealing approach detailed above gives us a good algorithm to find a global minimum:

1. Start with $\beta$ small (i.e. high temperature regime): the algorithm will then visit all the states of $S$ quite uniformly at the beginning. After a sufficiently high number of iterations, the Metropolized chain is roughly distributed as $\pi_\beta$.
2. Increase then $\beta$ (i.e. lower the temperature) and rerun the algorithm from the state found in the previous step.
3. Repeat step 2 until $\beta$ is sufficiently large, so that one can be quite sure (i.e. with prob. $1 - \epsilon$) to have reached a global minimum.